# Device and Method for Synthesizing Speech

## CROSS REFERENCE TO RELATED APPLICATIONS

All the content disclosed in Japanese Patent Application No. H11-285125 (filed on October 6, 1999), including specification, claims, drawings and abstract and summary is incorporated herein by reference in its entirety.

## BACKGROUND OF THE INVENTION

1. Field of the invention

This invention relates to speech processing like speech synthesis and, more particularly, to pitch conversion process.

2. Description of the related art

Concatenative Synthesis is a known speech synthesis. In this method, speech sound is synthesized by means of concatenating the prepared sound waveforms. However, there is a problem that natural sounding speech can not be obtained simply from the concatenation of the prepared waveforms because of the incapability of intonation control.

In order to solve this problem, PSOLA (Pitch Synchronous Overlap Add) method has been suggested. In this method, speech sound with the different pitch length can be obtained by filtering two pitch-unit speech waveforms through a Hanning window and making them slightly overlapped each other. (E. Moulines et. al, "Pitch-Synchronous waveform processing techniques for text-to-speech synthesis using diphones" Speech Communication, 1990.9)

Referring to Fig. 22 and Fig. 23, the PSOLA method is described as follows. Fig. 22 shows a part of speech waveform. The waveform is repeated almost periodically. This one repeating unit is a pitch. Pitch of the

sound varies depending on this pitch length.

In the PSOLA method, at first, a waveform is clipped out with its peak point of M as a center using a Hanning window as shown in Fig. 23. Next, the clipped waveforms are overlapped until their pitch lengths agree with the target pitch length. The width of the Hanning window for filtering is set in such a way that the clipped waveforms will be overlapped by one half. Thus, pitch can be converted to minimize the generation of undesirable frequency components. Therefore, if pitch is converted by modifying fundeamental frequency using the PSOLA method, the intonation can be controlled.

However, the PSOLA method still has following problems.

Firstly, as shown in Fig. 24 to Fig. 27, unnatural reduction of amplitude might happen in the segment where waveforms are overlapping. Fig. 24 shows an original waveform (indicated with a damped sine wave for easy understanding). Fig. 25 shows the waveform filtered through the left side components of a Hanning window. Fig. 26 shows the waveform filtered through the right side components of a Hanning window. Fig. 27 shows a composite waveform. As indicated in Fig. 27, the unnatural reduction in amplitude appears in the middle part of a pitch. This amplitude reduction causes a distortion of microstructure of speech waveform represented by formant.

Secondly, another problem is that echoes are produced with the contiguous pitch peaks as shown in Fig. 28. This is indicated in H. Kawai, et.al. "A study of a text-to-speech system based on waveform splicing," Tech. Rep. of the Institute of Electronics, Information and Communication Engineers, SP93-9, pp.49-54, Japan (1993,5) (in Japanese, the abstract in English). In this literature, the writer proposes the use of a trapezoidal window. However, using the mentioned trapezoidal window

might still produce undesirable frequency components during the process of overlapping that make the synthesized sound unnatural.

As shown in Fig. 1, a speech waveform in a pitch-unit is considered to be divided into two segments: 1) the segment of $\beta$, that starts from the minus peak at which the waveform depending on the shape of vocal tracts appears and 2) the segment of $\gamma$ at where the waveform, depending on the vocal tract shape, is attenuating and converging on the next minus peak. In addition, $\alpha$ in Fig. 1 is the point at which a minus peak appears along with the glottal closure. In the described PSOLA method, the center of the Hanning window is set at around the peak of M during a pitch with the goal of maintaining the contour of waveform around the peak of M. However, putting too much emphasis on the maintenance of the waveform contour around the peak brought about the above-described problems.



## SUMMARY OF THE INVENTION

It is an object of the present invention to provide a pitch conversion process technology capable of solving the problems described above and of minimizing the distortion of the naturalness of speech sound.

In order to achieve this object, the present invention processes waveform by converting pitch in the segment of $\gamma$ just before the next minus peak, which is least affected by the minus peak associated with the glottal closure, on the basis of the described characteristics of speech waveforms. As such, waveform processing can be performed by keeping the complete contour of waveform at around the peak and thereby reducing the effects of pitch conversion.

Moreover, according to the present invention, the sampled speech waveforms to find out which part of pitch is consistent. Fig. 2 shows several pitch-unit waveforms of /a/. It is apparent that waveforms are similar until 2.5

ms. From that point, they stay at around zero value, and, from a certain point, they simply decline and converge on the minus peak value. Therefore, it is clear that the pitch difference of each waveform in actual utterances is dependent upon the difference in duration of a zero value segment or the difference of the start point of a simple declining segment. Consequently, it has been found out that an optimal pitch conversion can be performed by processing the segment of $\gamma$ and, particularly, the zero value area.

In accordance with characteristics of the present invention, there is provided a speech synthesis device comprising:

speech database storing means for storing sample waveform data in a speech unit and a speech database created by associating the sample sound waveform data with their corresponding phonetic information;

speech waveform composing means for dividing phonetic information into speech units upon receiving the phonetic information of speech sound to be synthesized, for obtaining sample speech waveform data corresponding to the each phonetic information in a speech unit from the speech database, and for generating speech waveform data to be composed by means of concatenating the sample speech waveform data in speech units; and

analog converting means for converting the speech waveform data received from the speech waveform composing means into analog signals;

wherein the speech waveform composing means comprises pitch converting means for converting pitch by means of processing a segment of a waveform in which the waveform is converging on a minus peak during a periodical unit of speech waveform data.

Also, in accordance with characteristics of the present invention, there is provided a computer-readable storing medium storing a program for executing pitch conversion using a computer, the program comprising the step of:

processing a segment of a waveform in which the waveform is

converging on a minus peak during a periodical unit of speech waveform data, upon receiving the speech waveform data requiring pitch conversion.

Further, in accordance with characteristics of the present invention, there is provided a speech synthesis device comprising:

speech database storing means for storing a speech database having several sample speech waveform data with various pitch lengths for each speech unit and phonetic information associated with these sample waveform data;

speech waveform composing means for dividing phonetic information into speech units upon receiving phonetic information of speech sound to be synthesized, for obtaining a desirable sample speech waveform data from among the sample speech waveform data corresponding to the divided phonetic information in a speech unit in the speech database, and for generating speech waveform data to be composed by means of concatenating the obtained sample speech waveform data in speech units; and

analog converting means for converting the speech waveform data received from the speech waveform composing means into analog signal;

wherein the speech database is constructed of several sample speech waveform data with various pitch lengths prepared by modifying a contour of a waveform in a segment in which the waveform is converging on the minus peak during a periodical unit of speech waveform data.

In accordance with characteristics of the present invention, there is provided a computer-readable storing medium storing a program for executing speech synthesis by means of a computer using a speech database, the program comprising the steps of:

receiving phonetic information of speech sound to be synthesized and dividing the phonetic information into speech units;

obtaining a desirable sample speech waveform data from among sample speech waveform data corresponding to the divided phonetic

information in a speech unit in the speech database; and

generating speech waveform data to be composed by means of concatenating the obtained sample speech waveform data in speech units;

wherein the speech database is constructed of several sample speech waveform data with various pitch lengths prepared by modifying a contour of a waveform in a segment in which the waveform is converging on a minus peak during a periodical unit of speech waveform data.

Also, in accordance with characteristics of the present invention, there is provided a computer-readable storing medium for storing several sample speech waveform data with various pitch lengths for each speech unit, wherein these several sample speech waveform data are prepared by modifying a contour of a waveform in a segment in which the waveform is converging on a minus peak during a periodical unit of speech waveform data.

Further, in accordance with characteristics of the present invention, there is provided a computer-readable storing medium for storing a speech database, the storing medium comprising:

a sample waveform data storing area storing sample waveform data of human speech utterances in a speech unit;

a phonetic information storing area storing phonetic information associated with the sample waveform data in the speech unit; and

an indicating information storing area that stores information to provide a last zero crossing point before a minus peak in the sample waveform data.

In accordance with characteristics of the present invention, there is provided a method of pitch conversion for speech waveform, the method comprising the step of:

performing pitch conversion by processing waveform in a segment in

which the waveform is converging on a minus peak during a periodical unit of speech waveforms.

Also, in accordance with characteristics of the present invention, there is provided a speech processing device for processing speech waveform in accordance with entered commands, wherein at least any one of amplitude, fundamental frequency or duration of speech is modified using corresponding icons or switches of the up arrow, the down arrow, the right arrow or the left arrow.

Further, in accordance with characteristics of the present invention, there is provided a computer-readable storing medium storing a program for implementing a speech processing device for processing speech waveform in accordance with entered commands, the program comprising the step of:

modifying at least any one of amplitude, fundamental frequency or duration of speech with using corresponding icons or switches of the up arrow, the down arrow, the right arrow or the left arrow using a computer.

Also, in accordance with characteristics of the present invention, there is provided a speech processing device for processing speech waveform in accordance with entered commands, wherein the up arrow at least to raise fundamental frequency and the down arrow at least to lower fundamental frequency are assigned.

Further, in accordance with characteristics of the present invention, there is provided a computer-readable storing medium storing a program for implementing a speech processing device for processing speech waveform in accordance with entered commands, the program comprising the step of:

assigning the up arrow at least to raise fundamental frequency and the down arrow at least to lower fundamental frequency using a computer.

In this invention, the term "speech unit" refers to a unit in which speech waveforms are handled, in speech synthesis or speech analysis.

The term "speech database" refers to a database in which at least speech waveforms and corresponding phonetic information are stored.

The term "speech waveform composing means" refers to means for generating a speech waveform associated with a given phonetic information according to rules or sample waveforms. In an embodiment of the present invention, steps S4 to S12 in Fig. 5 and Fig. 6 are corresponding to this speech waveform composing means.

The term "periodical unit" refers to a unit of speech waveform that repeats periodically. In an embodiment of the present invention, pitch is corresponding to a periodical unit.

The term "arrow" refers to a sign indicating or suggesting a direction including, for example, a triangle as a direction indicator.

The term "storing medium on which programs or data are stored" refers to a storing medium including, for example, a ROM, a RAM, a flexible disk, a CD-ROM, a memory card or a hard disk on which programs or data are stored. It also includes a communication medium like a telephone line and other communication networks. In other words, this term includes not only the storing medium, like a hard disk which stores programs executable directly upon connection with CPU, but also the storing medium like a CD-ROM etc., which stores programs executable after being installed in a hard disk.

Further, the term "programs (or data)" here, includes not only directly executable programs, but also source programs, compressed programs (or data), and encrypted programs (or data).

Other objects and features of the present invention will be more apparent to those skilled in the art on consideration of the accompanying drawings and following specification, in which are disclosed several exemplary embodiments of the present invention. It should be understood that variations, modifications and elimination of parts may be made therein as fall within the scope of the appended claims without departing from the spirit of the invention.

## BRIEF DESCRITION OF THE DRAWINGS

Fig. 1 is a graph showing a part of the speech waveform of /a/;

Fig. 2 is a graph showing many waveforms of /a/ overlapping one another;

Fig. 3 is a diagram illustrating an overall configuration of the speech synthesis device according to a representative embodiment of the present invention;

Fig. 4 is a block diagram illustrating a hardware configuration of the device shown in Fig. 3;

Fig. 5 is a flow chart showing the speech synthesis processing program;

Fig. 6 is a flow chart showing the speech synthesis processing program;

Fig. 7 is a flow chart showing the program for pitch conversion processing;

Fig. 8 is a table illustrating the contents of a word dictionary;

Fig. 9 is a table illustrating the contents of a dictionary of syllable duration;

Fig.10 is a view showing a part of the analysis table;

Fig.11 is a graph showing the determined contour of fundamental frequency;

Fig.12 is a table illustrating the contents of a dictionary of voiced/unvoiced sounds of consonants/vowels;

Fig.13 is a table illustrating the contents of a dictionary of sound source

amplitude ;

Fig.14 is a view showing the contents of a speech database;

Fig.15 is a graph showing pitch modification with zero value insertion;

Fig.16 is a graph showing pitch shortening with other than zero value deletion;

Fig.17 is a table illustrating the definitions of Extended CV;

Fig.18 is a diagram illustrating an overall configuration of the second embodiment of the present invention;

Fig.19 is a graph showing the contents of a speech database;

Fig.20 is a view illustrating icons for operation;

Fig.21 is a flow chart showing the program for judging entered commands;

Fig.22 is a graph showing pitches of speech sound;

Fig.23 is a view illustrating pitch conversion process by using PSOLA method;

Fig.24 is a graph showing the effect of processing using PSOLA method (original waveform);

Fig.25 is a graph showing the effect of processing using PSOLA method (left side components of Hanning window);

Fig.26 is a graph showing the effect of processing using PSOLA method (right side components of Hanning window);

Fig.27 is a graph showing the effect of processing using PSOLA method (composite waveform);

Fig.28 is a schematic illustration of echo generation by using PSOLA method.

## DETAILED DESCRIPTION OF REPRESENTATIVE EMBODIMENTS

### 1. THE FIRST EMBODIMENT

#### (1) Overall Structure

Fig. 3 shows an overall structure of the speech synthesis device according to the first representative embodiment of the present invention. In

this embodiment, speech waveform composing means 16 comprises character string analyzing means 2, speech unit obtaining means 4, waveform converting means 12, and waveform concatenating means 22. Moreover, the waveform converting means 12 comprises duration converting means 6, amplitude converting means 8 and pitch converting means 10.

A provided character string is morphologically analyzed with the character string analyzing means 2, referring to a dictionary for morphological analysis 20. The character string is divided into speech units. Further, character string analyzing means 2, by referring to the environment of the preceding and succeeding sequences of sounds, determines the voiced and unvoiced sounds classification, duration, the contour of amplitude, and the contour of fundamental frequency for each speech unit by referring to the dictionary for morphological analysis 20.

Upon receiving the result of morphological analysis from character string analyzing means 2, speech unit obtaining means 4 obtains sample speech waveforms in each speech unit from a speech database 18. The duration converting means 6 converts the duration of the obtained sample speech waveforms in accordance with the duration provided by the character string analyzing means 2. And amplitude converting means 8 converts the amplitude of the obtained sample speech waveforms in accordance with the amplitude provided by the character string analyzing means 2. The pitch converting means 10, in accordance with the contour of fundamental frequency provided by the character string analyzing means 2, converts the pitch of the obtained sample speech waveforms. The sample speech waveforms in each speech unit, as desirably processed as described above, are concatenated by means of the waveform concatenating means 22. Thus, a speech waveform data is produced.

Analog converting means 14 converts this speech waveform data into

analog sound signals and produces output.

(2) Hardware Configuration

Fig. 4 shows an embodiment of a hardware configuration using a CPU for the speech synthesis device of Fig. 3. Connected to a CPU 30 are a memory 32, a keyboard/mouse 34, a floppy disk drive (FDD) 36, a CD-ROM drive 40, a hard disk 44, a sound card 54 forming the analog converting means, and a display 58. Stored in the hard disk 44 are an operating system (OS) 52 such as WINDOWS 98™ by Microsoft™, a speech synthesis program 46, a speech database 48 and a dictionary for morphological analysis 50. These programs are installed from the CD-ROM 42 using the CD-ROM drive 40.

In this embodiment, the speech synthesis program 46 performs its functions in combination with the operating system (OS) 52. However, the speech synthesis program 46 may perform a part of or all of its functions by itself.

(3) Speech Synthesis Processing

Fig. 5 is a flow chart showing the speech synthesis program stored in the hard disk 44. First, an operator inputs a character string corresponding to the speech sound to be synthesized, using the keyboard 34 (step S1). Alternatively, a character string stored on the floppy disk 38 or transferred from other computers through networks may be used.

Next, the CPU 30 performs morphological analysis of the character string using reference to the word dictionary in the dictionary for morphological analysis 50 (step 2). The contents of this word dictionary are shown in Fig. 8. Then, the CPU 30, using reference to the word dictionary, breaks up the character string into words and obtains the pronunciation of each word. For example, when the character string of "ko n ni chi wa" is provided, a pronunciation as "/koNnichiwa/" is obtained.

Furthermore, accent value of syllables constituting a word is obtained for each word (step S3). Consequently, syllables of "ko" "N" "ni" "chi" "wa" together with their accent value as shown in Fig. 8 are obtained. Accent value depends upon the environment of the preceding and succeeding sequences of sounds. Therefore, the CPU 30 modifies the accent value using rules based on the relationships with the preceding and succeeding sequence of phonemes or syllables.

All syllables and their duration shown in Fig. 9, are stored in a dictionary of syllable duration in the dictionary for morphological analysis 50 on the hard disk 44. The CPU 30 obtains the syllable duration for each syllable by referring to the dictionary of syllable duration. Further, the CPU 30 modifies the duration based on the relationships with the preceding and succeeding sequence of phonemes or syllables (step S4 of Fig. 5)). Thus, a table for each syllable is prepared, as shown in Fig. 10.

As shown in Fig. 12, all phonemes and their classification of voiced/unvoiced sound are stored in a dictionary of voiced/unvoiced sounds of consonants/vowels in the dictionary for morphological analysis 50. In the index column in Fig. 12, "V" denotes vowels (voiced sound), "CU" denotes unvoiced sound of consonants and "CV" denotes voiced sound of consonants. The CPU 30 makes a voiced/unvoiced classification for each phoneme of "k" "o" "N" "n" "i" "ch" "i" "w" "a" by referencing to this dictionary. Furthermore, the CPU 30 determines voiced sounds that are uttered unvoiced by reference to a devoicing rule. Thus, each phoneme is classified into voiced or unvoiced sound (step S5 of Fig. 5).

Next, the CPU 30 generates the contour of fundamental frequency Fo as shown in Fig. 11, according to the table in Fig. 10 (particularly to the accent value) (step S6 of Fig. 5). In the unvoiced segments, the fundamental

frequency is not appearing.

Next, the contours of voiced sound source amplitude Av and unvoiced sound source amplitude Af are determined (step S7 of Fig. 5). In a dictionary of sound source amplitude in the dictionary for morphological analysis 50, the contours of sound source amplitude corresponding to each syllable are stored as shown in Fig. 13. The CPU 30, referring to this dictionary, determines voiced sound source amplitude Av and unvoiced sound source amplitude Af for each syllable of "ko" "N" "ni" "chi" "wa". In addition, the CPU 30 modifies the obtained sound source amplitude according to the accent value and the environment of the preceding and succeeding sequences of sounds. Moreover, the CPU 30 modifies the contour of sound source amplitude to conform to the determined syllable duration in step S4 in Fig. 5.

Then, the CPU 30 obtains the sample speech waveforms for each syllable from the speech database 48. As shown in Fig. 14, the speech database 48 stores sample speech waveforms of real speech utterance that are divided into syllables and accompanied by phonetic information. Moreover, the contour of sound source amplitude, the contour of fundamental frequency, duration, a pitch mark and a zero crossing mark for each syllable are also stored in the speech database 48. The pitch mark here refers to a mark assigned to the location of the peak value at each pitch unit (see M in Fig. 1). The zero crossing mark refers to a mark assigned to the last zero crossing point before the minus peak for each pitch unit (see $\alpha$ in Fig. 1). In this embodiment, the pitch marks and the zero crossing marks are given with a time.

Because a massive number of sample waveforms are stored in the speech database, there is more than one sample waveform corresponding to one syllable, for example "ko". Therefore, the CPU 30 searches and obtains the optimal sample waveform for each syllable with reference to the relation

with the preceding and succeeding syllable sequences (step S8 in Fig. 5).

Next, the CPU 30 modifies the sample speech waveform for each syllable so that the duration of the sample speech waveform obtained from the speech database 48 may conform to the duration determined in step S4 of Fig. 5 (step S9 in Fig. 6). This modification is made by duplicating (inserting the same waveforms) or deleting a few pitch-unit waveforms.

Then, the CPU 30 modifies the sample speech waveform obtained from the speech database 48 for each syllable so that its contour of amplitude may conform to the contour of amplitude determined in step S7of Fig. 5 (step S10 in Fig. 6).

Furthermore, the CPU 30 modifies the sample speech waveform obtained from the speech database 48 for each syllable so that its contour of fundamental frequency may conform to the contour of fundamental frequency determined in step S6 of Fig. 5 (step S11 in Fig. 6).

Fig. 7 is a flow chart showing the program for pitch conversion processing. Pitch conversion processing is performed only for the waveform of voiced sounds, because the waveform of unvoiced sounds does not contain the regular periodical repeats.

First, the CPU 30 obtains the fundamental frequency of the first pitch-unit of the sample speech waveform for the target syllable, from the contour of fundamental frequency data in the speech database 48. Next, the CPU 30 obtains the corresponding fundamental frequency using reference to the contour of fundamental frequency determined in step S6 of Fig. 5. Then the CPU 30 determines whether or not both fundamental frequencies are matching in step S22. If they are matching, the process goes to the next step S26(Fig. 7) since no pitch conversion is required.

If, in step S22, the CPU 30 determines that the fundamental frequencies do not much, then in step S23 of Fig. 7., the CPU 30 determined whether the pitch of sample sound waveform shall be lengthened (lowering fundamental frequency) or shall be shortened (raising fundamental frequency). Resulting from this judgement, pitch is lengthened (step S25) or shortened (step S24).

The CPU 30 finds out the last zero crossing point right before the minus peak in the objective pitch. The zero crossing point is easily determined because it is stored on the speech database as shown in Fig. 14.

To lengthen pitch, a zero value segment is inserted at this zero crossing point as shown in Fig. 15.

On the contrary, to shorten pitch, in case that there is an almost zero value segment around the zero crossing point, the segment is to be deleted as needed. In case that there is not a zero value segment at around the zero crossing point, the following operation as shown in Fig. 16 is performed to shorten pitch (shortened duration is N). Firstly, the frame between 2N-1 and N before the minus peak is filtered through the Hanning window with window magnitude of 1 at 2N-1 and 0 at N. Likewise, the frame before the minus peak between N-1 and the minus peak is filtered through the Hanning window with window magnitude of 1 at the minus peak and 0 at N-1 before the minus peak. The merger of waveform elements derived from the two window filtering is adopted as a modified waveform. Thus, a 2N frame is shortened to an N frame.

Alternatively, the above window processing may be performed by setting the window magnitude of 0 at around the location of zero crossing. The farther processing point is from zero crossing, the larger magnitude up to

1 is applied. Thus, at the point far from zero crossing point, the window magnitude of 1 is applied so that the waveform may be kept as it is, and the window magnitude of 0 is applied at the zero crossing so that the waveform may be substantially deleted. Accordingly, pitch can be shortened to minimize the distortion of naturalness by means of applying a bigger processing value to the segment around zero crossing, which is considered to be less influenced due to smaller amplitude.

After processing the pitch, the CPU 30 determines whether all pitch-unit waveforms have been likewise processed (step S26 of Fig.7). If not all pitch-unit waveforms have been processed, then the steps from S22 (Fig. 7) forward are repeated for the non-processed pitch in the next step (S27 of Fig. 7. After processing all pitch-unit waveforms, the pitch conversion processing for the objective syllable is completed. A fine adjustment in duration is made when it is required after pitch conversion. The pitch conversion processing is carried out for all syllables in the selected sample waveform.

After completing the pitch conversion processing as described above, the process goes to the step S12 in Fig. 6. In the step of S12, the composed speech waveform is obtained by way of concatenating the sample waveform modified for each syllable. Finally, the CPU 30 provides this composed speech waveform to the sound card 54. The sound card 54 converts this waveform into analog signals and produces output through the speaker 56.

(4) Other Embodiment of Speech Database

In the embodiment described above, the speech database (speech corpus) stores a large number of sample waveforms, assuming each syllable to be a speech unit. However, the present invention may also use a database that stores sample waveforms under the assumption that a phoneme is a speech unit. Or, in case that there is a contiguous sequence of more than one syllable without clear distinction, these syllables, in addition to one syllable,

may be treated as one cluster of syllables (Extended CV). Its definition is described in Fig. 17. A heavy syllable is given a higher priority than a light syllable and a superheavy syllable takes precedence over a heavy syllable, when they are extracted from the speech database. For instance, if a sequence of syllables is regarded as a superheavy syllable, a part of the sequence is not cut apart and extracted as a heavy syllable. Likewise, if a sequence of syllables is regarded as a heavy syllable, a part of the sequence is not cut apart and extracted as a light syllable. Accordingly, by treating a contiguous sequence of more than one syllable without clear distinction as one speech unit", continuity distortion can be avoided. In a representative embodiment of the present invention, employing at least a light syllable and a heavy syllable is recommended.

The speech corpus is used in the embodiment described above. However, the speech database that stores one speech waveform data each per one syllable (one phoneme or one Extended CV) may be used. Furthermore, the speech database storing one pitch-unit waveform data each per one syllable (one phoneme or one Extended CV) may also be used.

Moreover, in the embodiment as described above, a zero-crossing mark is stored on the speech database. However, a zero-crossing mark may be searched for at every time of processing in accordance with a pitch mark and so on, instead of being pre-stored on the speech database.

(5) Other Embodiment of Pitch Conversion Processing

In the embodiment described above, pitch change is performed by means of inserting or deleting a substantial zero value segment at zero crossing point. However, pitch may be changed by means of time compression or time extension of the segment where the waveform is declining and converging on the minus peak (see γ in Fig. 1). Generally, the time compression and time extension might generate undesirable frequency

components that have no relation with pitch conversion. However, since the waveform is simply declining and does not contain many frequency components in the segment of γ, distortion to speech sound quality is considered small.

By the way, instead of carrying out the processing of time compression or time extension of the segment of γ evenly, the intensive time processing may be performed at around zero crossing, and the farther from the zero crossing the less intensive time processing may be performed.

## 2. THE SECOND EMBODIMENT OF THIS INVENTION

Fig. 18 shows an overall configuration of the speech synthesis device of the second embodiment of the present invention. In this embodiment, speech waveform composing means 16 comprises character string analyzing means 2, speech unit waveform generating means 90, and waveform concatenating means 22. For generating a speech unit (such as a syllable), a speech database 18 stores several pitch-unit waveforms of speech sound, which are slightly different from one another in pitch. For instance, many pitch-unit waveforms for generating a syllable of /a/ are stored in various pitch lengths slightly different by about 1 ms. All other syllables (voiced sounds) are stored in a similar manner. For unvoiced sounds, noise waveforms are stored on the speech database 18.

A provided character string is morphologically analyzed with the character string analyzing means 2, referring to a dictionary for morphological analysis 20. The character string is divided into speech units. In addition, with referring to the environment of the preceding and succeeding sound sequence, the voiced and unvoiced sounds classification, the duration, the contour of amplitude, and the contour of fundamental frequency are determined for each speech unit by referring to the dictionary for morphological analysis 20.

The speech unit waveform generating means 90 obtains a pitch-unit waveform required for generating each speech unit, from the speech database 18. On this occasion, the waveforms with proper pitch length at each time are selected and picked out in accordance with the contour of fundamental frequency provided by the character string analyzing means 2. Then, the speech unit waveform generating means 90 modifies these pitch-unit waveforms with reference to the duration and the contour of amplitude provided by the character string analyzing means 2, and generates a waveform in a speech unit by means of concatenation. As for unvoiced sounds, the speech unit waveform generating means 90 generates waveforms using reference to the noise waveforms.

The speech waveforms in each speech unit generated as described above are concatenated with the waveform concatenating means 22. Thus, a speech waveform data is produced.

The analog converting means 14 converts this speech waveform data into analog sound signals and produces output.

Fig. 4 shows a representative embodiment of a hardware configuration using a CPU for the speech synthesis device of Fig. 18. In this embodiment, a waveform in a speech unit (such as a syllable) is synthesized by means of concatenating pitch-unit waveforms. For this reason, a lot of speech sound data of pitch-unit waveforms with various pitch lengths are prepared in the database 18 as shown in Fig. 19. These pitch length variations are obtained through the insertion of zero value segments at the last zero crossing point just before the minus peak.

In this embodiment as in the previous embodiment, pitch conversion may be performed at every time of processing. In this manner, there is no

need to prepare the various pitch length data. Instead, only one kind of pitch length data must be stored in the speech database 18.

In addition, as one of ordinary skill in the art would appreciate, the other embodiments described in the first embodiment may be applied to this second embodiment.

3. OTHER EMBODIMENT

In the above embodiments, pitch conversion is processed in accordance with the result of the analysis carried out by means of the character string analyzing means 2. However, the pitch conversion may be performed through commands entered by an operator.

Fig. 20 shows an example of a screen display for entering these commands. Fig. 21 is a flow chart showing the program for judging entered commands stored on the hard disk 44.

Amplitude and fundamental frequency of speech sound are raised by clicking the icon 200 (up arrow) with the mouse 34 (steps S50 and S53). In the same way, amplitude and fundamental frequency of speech sound are lowered by clicking the icon 204 (down arrow) (steps S50 and S52 of Fig, 21). Duration of speech sound is shortened by means of, for instance, deleting several pitch-unit waveforms, by clicking the icon 206 (left arrow) (steps S50 and S51 of Fig. 21). On the other hand, duration of speech sound is lengthened by means of, for instance, duplicating several pitch-unit waveforms, by clicking the icon 202 (right arrow) (steps S50 and S54 of Fig. 21).

While the methods of pitch modification described so far are preferable, other method may be also applied.

Thus, a pair of arrows (the up arrow and the down arrow, or the left arrow and the right arrow) corresponds to the processing of two opposite modifications. Accordingly, the contents of processing are intuitionally understandable, providing an easy operation for entering commands.

Alternatively, instead of using the screen icons in the above embodiment, entry switches shaped as an arrow or with an indication of an arrow may be used.

In the above embodiment, each of the up arrow and the down arrow corresponds to the processing of both amplitude and fundamental frequency. However, the processing of each one, two, or three of amplitude, fundamental frequency, and utterance duration may be arranged to be assigned to each arrow. This arrangement applies also to the right arrow and the left arrow. Furthermore, obliquely pointing arrows may also be adopted and assigned for both tasks associated with vertically pointing arrows and horizontally pointing arrows.



## 4. OTHER ASPECTS OF THE PRESENT INVENTION

While, in the above embodiment, a CPU is used to provide the respective functions shown in Fig. 3 and Fig. 18, a part or all of the functions may be given by using hardware logic.

The speech synthesis device according to claims 1 and 3 is characterized by comprising pitch converting means for converting pitch by means of processing a segment of a waveform in which the waveform is converging on a minus peak during a periodical unit of speech waveform data.

Therefore, the waveform can be processed in the segment that is less affected by the minus peak associated with the glottal closure, and then pitch can be converted without distorting naturalness.

The speech synthesis device according to claim 6 is characterized by providing the largest processing value at around zero crossing point and the smaller value at the farther from zero crossing point, within the segment in which waveform is converging on the minus peak.

Accordingly, pitch can be adjusted without distorting naturalness since the processing complies with the actual speech sound characteristic that each duration of zero value segment is different.

The speech synthesis device according to claim 7 is characterized by shortening or lengthening pitch by means of compressing or extending waveform along the time axis in the segment in which the waveform is converging on the minus peak.

Consequently, waveform can be processed through time compression or time extension in the segment that is less affected by the minus peak associated with the glottal closure. As such, pitch can be converted without a distortion in naturalness.

The speech synthesis device according to claim 8 is characterized by performing waveform processing at around zero crossing point within the segment where the waveform is converging on the minus peak. Therefore, processing can be performed in the segment that is less affected due to rather small amplitude.

The speech synthesis device of claim 9 is characterized by performing waveform processing at around zero crossing point by means of either inserting a substantial zero value segment to lengthen pitch or of eliminating a substantial zero value segment to shorten pitch.

Therefore, pitch can be converted, minimizing the influence on a spectrum. In addition, a simple operation like insertion and deletion of a zero value segment makes the waveform processing speedy.

The pitch converting method for speech waveform according to claim 11 is characterized in that pitch conversion is performed by way of processing waveform in the segment in which the waveform is converging on the minus peak during the periodical unit of speech waveforms.

Consequently, waveform can be processed in the segment that is less affected by the minus peak associated with the glottal closure, and pitch can be converted without losing naturalness.

The speech processing device according to claim 12 is characterized by modifying at least any one of amplitude, fundamental frequency or duration of speech with using corresponding icons or switches of the up arrow, the down arrow, the right arrow, or the left arrow.

Accordingly, speech sound amplitude, fundamental frequency or duration can be converted in a simple operation.

The speech processing device according to claim 14 is characterized by assigning the up arrow at least to raise fundamental frequency and the down arrow at least to lower fundamental frequency. Therefore, the present invention provides an easy-to-use, intuitive user interface for pitch conversion processing.

While the embodiments of the present invention, as disclosed herein, constitute preferred forms, it is to be understood that each term was used as illustrative and not restrictive, and can be changed within the scope of the claims without departing from the scope and spirit of the invention.